Amendments to the Specification

Please replace the first paragraph on page 5 with the following:

A summary of the operating conditions for multi-level storage is given in Figure ~~3B~~ 7C. During read, the following conditions need to be met: the voltage of the unselected control gate within a selected memory cell must be greater than the threshold voltage of the control + source voltage. The word select gate in the control gate pair is raised to the threshold voltage of the word gate + an override delta of around 0.5V + source voltage (Vt-wl + Voverdrive +Vs). Un-selected MONOS cells will be disabled by reducing the associated control gates to 0V. Program conditions are: Word line voltage is greater than threshold + an overdrive voltage delta for low current program. Both control gates in the selected pair are greater than Vt-high (the highest threshold voltage within the range of multi-level thresholds) + override delta. Adjacent memory cells sharing the same word line voltage are disabled by adjusting the control gates only.

Please amend paragraphs 6 and 7 on page 6 as follows:

Figs. ~~5B, 5C, and 5F~~5A through 5C are cross sectional representations of a second preferred embodiment of the process of the present invention.

Figs. 6A through ~~6F~~ 6E are cross sectional representations of a third preferred embodiment of the process of the present invention.

Please replace the last full paragraph on page 10 through the last paragraph on page 12 with the following:

The preceding processes describe fabrication of planar channel floating gates with very short channel (30 to 50nm). By modifying and adding a few process steps, a step split structure with more efficient ballistic injection can be fabricated using the same process integration scheme as the planar structure. This second embodiment of the present invention will be described with reference to Figs. ~~5B, 5C, and 5F~~5A through 5C.

After forming disposable sidewall spacer 242 by etching vertically the doped polysilicon, the silicon oxide layer 221 is vertically etched which corresponds to Fig 4B. In order to form a step split memory cell, the deviation starts at this point by continuing to etch into the silicon substrate by approximately 20 to 50nm. Then the bottom of the step is lightly implanted with Arsenic to form N-region 203 using the poly sidewall as a mask as shown in Fig ~~5B~~5A, where the dosage is about 3E13 to 4E13/cm^2 at 10 to 15KeV. Next, the N+ doped polysilicon disposable spacer is selectively removed by a wet etch (HNO_3/HF/Acitic acid, or H_3PO_4 or NH_4OH) or a dry plasma etch to the lightly doped bulk N- region. The bulk etching during this disposable spacer etch can be included as part of step etching. After gently etching off the left over gate oxide 221 under the disposable polysilicon spacer, the silicon surface is cleaned. The total step into silicon should be about 20 to 50 nm. If the step corner is sharp, corner rounding by rapid thermal anneal (RTA) at between about 1000 to 1100° C for about 60 seconds can be added as an option or a hydrogen anneal at 900°C and at a pressure of 200 to 300 mtorr can be performed. After these modifications and additions, the fabrication sequence returns to the procedures described previously.

Referring to Fig. ~~5C~~5B, a composite layer of oxide-nitride-oxide is formed. Layer 230 is shown without the three layers for simplicity. The bottom oxide is thermally grown and the thickness is between 3.6 and 5 nm, which is slightly thicker than the limit of direct tunneling (3.6nm), the silicon nitride layer deposited by chemical vapor deposition (CVD) is about 2 to 5 nm, and the top oxide is deposited by CVD deposition and is between about 4 and 8nm. Thermal oxidation may be added to improve the top oxide quality. Also, short nitridation in an N_2O environment can be added to improve the bottom oxide reliability prior to the deposition of the nitride layer.

Then an insitu phosphorous-doped polysilicon layer, which becomes the control gate, is deposited having a thickness of between 90 to 180 nm, and a vertical or anisotropic polysilicon etch is performed to form the sidewall gate 240, a shown in Fig ~~5C~~5B. By following the process steps given for the planar split device, the step-split device can be fabricated as shown in Fig ~~5F~~5C. This sidewall polysilicon gate can be silicided or replaced by refractory silicide as utilized in the first embodiment of the flat channel MONOS twin cell.

In the above process steps for both the planar and step devices, the disposable side wall spacer 242 can be plasma nitride or oxynitride or Boron Phosphorus Silicate Glass (BPSG) instead of polysilicon, since the etching rate of that material to the thermal silicon oxide can be very high (for example at least 10-100 times) in H_3PO_4 acid or diluted HF.

A third embodiment of the present invention will be described with reference to Figs. 6A ~~6D~~ 6E ~~and 6F~~. The third embodiment of the present invention will be a simplified process of the first embodiment of the planar twin MONOS memory cell with a slight program

speed penalty because controllability will be lost due to the usage of a single large spacer instead of two side wall spacers. Deviation from the normal CMOS process starts prior to deposition of word gate polysilicon 245. A composite layer of oxide-nitride-oxide (ONO), 230 in Fig 6A, is formed. Layer 230 is again shown without the three layers for simplicity. The bottom silicon oxide layer is preferred to be grown thermally with a thickness of between about 3.6nm to 5nm, the silicon nitride layer deposited by CVD deposition is about 2 to 5 nm and the top oxide layer is deposited by CVD deposition and about 5 to 8 nm thick. The top oxide CVD layer is slightly thicker compared to the first and second process embodiments, for subsequent polysilicon and disposable sidewall spacer etch stop. Then the polysilicon 245 for gate material is deposited by CVD and followed by CVD silicon nitride 232 deposition thickness of between about 50 to 100 nm.

Please replace the third paragraph on page 14 with the following:

Then the nitride layer 232 is selectively etched by H_3PO_4 or etched by a chemical dry etch. The polysilicon layer 248 having a thickness of between 150 and 200 nm is deposited by CVD. This polysilicon layer and underlying word gate polysilicon 245 are defined by normal photoresist and RIE processes. The structure at this point is as shown in Fig ~~6F~~6E.

Please replace the first paragraph on page 18 with the following:

During read operation of nitride region 313, shown in Fig. ~~3C~~7C, the source line 321 can be set to some intermediate voltage (~1.2V) and the bit line 322 may be precharged to 0V. In addition, the following conditions must be met in order to read a selected nitride charge region: 1) the word select gate voltage must be raised from 0V to a

voltage (2.5V) which is some delta greater than the sum of the threshold voltage of the word select gate (Vt-wl=0.5V) and the source voltage (1.2V), and 2) the voltage of the control gate above the selected nitride charge region must be near Vt-hi ("express"). The voltage of the control gate above the unselected nitride charge regions must be greater than the source voltage plus Vt-hi ("over-ride"). The control gates above the unselected adjacent nitride charge regions must be zero ("suppress"). The voltage of the bit diffusion 322 can be monitored by a sense amplifier and compared to a switch-able reference voltage, or several sense amplifiers each with a different reference voltage, to determine the binary value that corresponds to nitride charge region 313's threshold voltage, in a serial or parallel read manner, respectively. Thus, by over-riding the unselected nitride region within the selected memory cell, and then suppressing the adjacent cell unselected nitride regions, the threshold state of an individual selected nitride region can be determined.

Please amend the Abstract as follows:

ABSTRACT

A fast low voltage ballistic program, ultra-short channel, ultra-high density, dual-bit multi-level flash memory is described. The structure and operation of this invention is enabled by a twin MONOS cell structure having an ultra-short control gate channel of less than 40nm, with ballistic injection which provides high electron injection efficiency and very fast program at low program voltages of 3~5V. The ballistic MONOS memory cell is arranged in the following array: each memory cell contains two nitride regions for one word

gate, and ½ a source diffusion and ½ a bit diffusion. Control gates can be defined separately or shared together over the same diffusion. Diffusions are shared between cells and run in parallel to the side wall control gates, and perpendicular to the word line. ~~The features of fast program, low voltage, ultra-high density, dual-bit, multi-level MONOS NVRAM of the present invention include: 1) Electron memory storage in nitride regions within an ONO layer underlying the control gates, 2) high density dual-bit cell in which there are two nitride memory storage elements per cell, 3) high density dual-bit cell can store multi-levels in each of the nitride regions, 4) low current program controlled by the word gate and control gate, 5) fast, low voltage program by ballistic injection utilizing the controllable ultra-short channel MONOS, and 6) side wall control poly gates to program and read multi-levels while masking out memory storage state effects of the unselected adjacent nitride regions and memory cells.~~